动 物 学 研 究 2000, Jun. 21 (3): 231～237 CN 53－1040/Q ISSN 0254－5853
Zoological Research

A NEW TECHNIQUE FOR PREDICTING DISTRIBUTION OF TERRESTRIAL VERTEBRATES USING INFERENTIAL MODELING

CHEN Guo-Jun[①] A. Townsend Peterson

(*Natural History Museum and Department of Ecology and Evolutionary Biology, The University of Kansas, Lawrence, KS 66045－2454, USA*)

Abstract: A new technique for predicting species' geographic distribution is described. The approach involves 3 steps: ①setting up geographic base data: ②collecting and georeferencing distributional points; ③modeling ecological niches using the biodiversity species workshop implementation of the genetic algorithm for rule-set prediction (GARP). To illustrate these procedures, an example based on the Brown Eared Pheasant (*Crossoptilon mantchuricum*) is developed. This technique constitutes a useful tool for assessing geographic distribution for questions of ecology, biogeography, systematics, and conservation biology.

Key words: Geographic information systems; Genetic algorithm for rule-set prediction; Distribution; Ecological niche

Geographic information system applications in conservation biology have proceeded in two general directions. The first consists of developing algorithms to predict species' geographic distribution. Various approaches are used to characterize species' ecological niches in multivariate environmental space; geographic areas fitting these conditions are then taken as predicted distribution, filling gaps created by uneven sampling (Nix, 1986; Walker *et al*., 1991; Carpenter *et al*., 1993; Sperduto *et al*., 1996). The second direction is prioritizing areas for conservation, as exemplified by Gap Analysis and other applications (e.g. Daniels *et al*., 1991; Russell-Smith *et al*., 1992; Bojorquez-Tapia *et al*., 1995; Harrison *et al*., 1995; Kiester *et al*., 1996), in which distributional information is integrated into strategies for reserve portfolio design. Integrated, these two efforts could provide a strong basis for educated decisions regarding geographic priorities for conservation (Peterson *et al*., 2000).

In China, a large-scale program entitled the "Terrestrial Vertebrate Wildlife Resources Survey" has been in process since 1996. This effort involves documenting distribution and population numbers for all endangered species, as well as for terrestrial vertebrate species of special interest economically or ecologically. A particularly difficult challenge has been determining species' distributional areas. New advances in geographic information systems and inferential computer software (e.g. Stockwell *et al*., 1991), however, offer important opportunities to overcome these challenges and take steps forward to understand better the distribution of Chinese terrestrial vertebrates.

In this paper, we describe an important advance in the first sector of geographic information system applications to conservation biology: how species' geographic distribution can be predicted using the Genetic Algorithm for Rule-set Prediction (GARP)

收稿日期：1999－08－26；修改稿收到日期：2000－01－15

基金项目：This research was supported by the National Science Foundation of the United States of America and The Ministry of Education of The People's Republic of China

[①] Address for correspondence: Guo-Jun CHEN
Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, Kansas 66045－2454, USA Tel.: (785)864－3095 Fax: (785)864－5335 E-mail: guojun@eagle.cc.ukans.edu

modeling system. We work out a concrete example based on an endangered species of pheasant, the Brown Eared Pheasant (*Crossoptilon mantchuricum*) endemic to China, and discuss potential applications of the approach.

1 Methods

The fundamental ecological niche of a species can be defined as the conjunction of ecological conditions within which it is able to maintain populations; as such, it is defined in multidimensional ecological/environmental space. The fundamental niche, the focus of modeling efforts, must be distinguished carefully from the realized niche (that part which is actually occupied), so as to maintain the modeling efforts focused on ecological dimensions important to a particular species.

1.1 Several approaches have been used to approximate species' fundamental ecological niches. The simplest is BIOCLIM (Nix, 1986), which involves tallying species' occurrences in categories of each environmental dimension, trimming marginal portions of distribution, and taking the niche as the conjunction of the trimmed ranges. Although easy to implement, and conceptually attractive, BIOCLIM suffers an odd reduction in efficacy when many environmental dimensions are included—numbers of environmental combinations simply overwhelm most sampling protocols. BIOCLIM also suffers in general from high commission error rates.

1.2 A second class of approaches is based on logistic multiple regression, techniques aimed at predicting probability of "yes" versus "no" in the dependent variable. This idea combines well with the concept of physiological tolerances determining species' presence along continuous climate dimensions, but does less well when categorical information (e.g. vegetation type and soil type) is also to be included. In effect, logistic regression divides environmental space into two portions ("habitable" and "uninhabitable"), an approach that may be useful under some circumstances. Implementations of this approach have included important improvements, such as relaxation of distributional assumptions regarding errors in the regression.

1.3 Finally, the Genetic Algorithm for Rule-set Prediction (GARP) includes both of the methods described above, as well as other set-based approaches in an iterative, artificial-intelligence-based approach (Stockwell *et al*., 1992). Here, individual algorithms are used to produce component "rules" in the broader rule-set, and hence portions of the species' distribution may be determined as within or without the niche based on different algorithms. As such, GARP should represent a superset of the other approaches, and should always have greater predictive ability than any one of them. Extensive testing of GARP has indicated excellent predictive ability[①] and insensitivity to BIOCLIM's problems with environmental data density (Peterson *et al*., 1999a).

The GARP algorithm works in an iterative process of rule selection, evaluation, testing, and incorporation or rejection. Occurrence data are divided into two halves: training data (for model building) and test data (for model evolution). First, a method is chosen from a set of possible tools [e.g. logistic regression, BIOCLIM rules (Nix, 1986), etc.], applied to the training data, and a rule developed. Predictive accuracy is evaluated based on 1 250 points resampled from the test data and 1 250 points sampled randomly from the study region as a whole, and accuracy calculated as the sum of points correctly predicted as present or absent, divided by the total number of points in the map (Stockwell *et al*., 1992). The change in predictive accuracy from one iterative to the next is used to evaluate whether a particular rule should be incorporated into the model. The algorithm runs 1 000 iterations ("generations") or until addition of rules has no appreciable effect on the accuracy measure ("fitness"). Complete details and documentation of the algorithm are available at http://biodi.sdsc.edu.

The principal steps in the modeling approach developed herein are assembly of base geographic data

[①] Ball L G, Peterson A T, Cohoon K C, Submitted. Predicting geographic distribution of tropical birds.

layers, collection of species occurrence records (longitude and latitude), building ecological niche models using GARP, and predicting geographic distribution. The following details each step.

1.3.1 Step one—Preparing base geographic coverages

Necessary for ecological niche modeling is an information base that includes environmental dimensions relevant to distributional and ecological limitation of the species in question. Important dimensions frequently include measures of temperature and precipitation (both averages and extremes), vegetation types, and elevation, among others. Additional environmental dimensions may be relevant to applications to particular taxa or regions. Geographic data can exist in the form of continuous or categorical information.

Many sources of these geographic coverages are available for building geographic base layers. An excellent source of varied information is ESRI's (1996) ArcAtlas, a set of maps of more than 40 geographic themes at scales of 1∶10 000 000 for Europe; 1∶20 000 000 for North and South America, Africa, and Antarctica; and 1∶25 000 000 for Asia and Australia. Other excellent sources of geographic base layers include the EROS data center for satellite imagery, digital elevation models, and other data types[①]. Many regional data sources are available for particular geographic applications.

All geographic base data layers are converted to raster grid format. Grids must be exactly coincident among coverages, including numbers of rows and columns, cell sizes, and grid locations. These operations can be achieved easily using the raster grid import/export capabilities of ArcView (versions 3 and later).

For Chinese applications, we have already developed a basic set of geographic base data coverages extending across all of China based on ESRI (1996). Included are coverages summarizing low and high temperature and precipitation in January, July, and year-round; low and high solar radiation; snow cover; geology; soils; geomorphology; and vegetation types. This set of base layers has been rasterized with pixes of 21.930 km × 16.997 km, and is available for public use on the BSW facility[②].

1.3.2 Step two—Distributional data for species

The second step in the process is aggregation sets of points representing known occurrences of the species in question. These data are available from a number of sources, including museum specimen tag data, monographic treatments that include locality information, and observational data sets (e.g. censuses, sightings, and compilations, as well as results of actual fieldwork). A particularly useful source, though in prototype stage, is the species analyst, a distributed data base including different biodiversity data bases worldwide[③]. Textual locality descriptions must be converted into standard latitude-longitude coordinates based on gazetteers or by reference to maps.

1.3.3 Step three—Modeling niches and predicting distribution

Once base geographic coverages have been prepared and mounted, and species' occurrence records assembled and georeferenced, analysis can begin. The BSW facility allows researchers to perform a variety of analyses on biodiversity data in real time on powerful computers at the San Diego Supercomputer Center via the World Wide Web. Latitude and longitude data are submitted from a web browser, and occurrence records can be mapped and manipulated, ecological niche models developed, and models projected as predicted geographic distribution. The BSW application is composed of four frames in a web browser, which are described and documented on the website. Most operations are menu-driven, making use convenient. A brief description of important commands and operations follows.

①Base data　The base data option allows selection of geographic base data for regions to be analyzed. On the BSW, world data sets are available at coarse scales, and regional data at finer scales. Data for China are available at a scale of resolution of pixels of

[①] http://edcwww.cr.usgs.gov/webglis

[②] http://biod.sdsc.edu

[③] http://habanero.nhm.ukans.edu

17 km × 22 km for analysis. Users are given options of selecting particular regions, or selecting from world data by limiting geographic coordinates.

②Biological data Here, two options are available. First ("upload"), species' occurrence points can be entered as a list of longitude-latitude pairs separated by spaces, one line per occurrence point. A blank line between lines of data plots sets of points in different colors. Points can be entered manually, pasted in from other applications, or imported as ASCII files. Alternatively ("from database"), for some taxa and regions, distributional data are provided for retrieval on the website, and can be accessed using this option.

③Modify data Here, particular geographic base layers and species' occurrence points can be visualized, and can be included or excluded from analysis. Additional occurrence points can be added interactively by clicking on the map image in the screen frame.

④Make model This frame is central to the BSW, presenting options for generating ecological niche models from which geographic predictions are developed. Four alternative prediction algorithms may be selected including BIOCLIM (an approach based on frequency ranges on environmental dimensions), E-ball (an approach based on distance measures), Logit (logistic multiple regression), and GARP. Herein, we focus on GARP, which has proven to be the best of the four methods in a variety of tests.

Several options are available for modifying and adjusting GARP models. The convergence criteria option is a parameter of the genetic algorithm in GARP; decreasing this parameter refines the requirement for stability of the final model, improving results, but also increasing the processing time. The resembling type option controls how training data are prepared: specifying 0 constitutes the training data in proportion to the frequency of values in the area, whereas specifying 1 populates the training data set in equal proportions of presences and absences (default). The frequency for dumping intermediate models option allows use of later options for reviewing the advance of the genetic algorithm by storing intermediate models.

⑤Model output The immediate output of the modeling procedure is in the form of an image map, which allows visualization of model results. Although these image maps are useful, and can readily be extracted into word-processing applications, several additional options allow users to enter deeper into the modeling process. The most useful of these options are as follows.

Accuracy allows model efficiency to be evaluated. Each rule is assigned an estimate of its predictive accuracy: its posterior probability. The dimension bar in this frame is a list of threshold probabilities for inclusion: specifying 0.8 presents accuracy based on rules with expected accuracy of 0.8 or better: specifying 0 evaluates all rules. Results are shown as a confusion matrix, with rows and columns representing the predicted and actual values, respectively. The cells on the diagonal are correct predictions, while cells off the diagonal represent incorrect predictions.

The alternatives option produces a list of rules in the model, together with associated performance indices. Predictions resulting from each rule can be viewed in isolation. Choosing among alternative rules is one way to allow experts to exercise their specialized expertise and biological insight into problems. The combine rules option allows users to display combinations of model rules singly or in combination for further exploration of data. Finally, the overlays option allows users to alter the form of output. Scale and extent of maps can be changed, data points added, and areas labeled. Perhaps most important are specifications for map download and output formats: as a.gif file for image viewing, in postscript format for Microsoft Word documents, and as ASCII raster grids for upload into GIS programs.

Once ecological niche models are in hand, additional possibilities open, including modeling distributional changes with climate change and outputs visualized in virtual reality.

2 Example

The brown eared pheasant (*Crossoptilon mantchuricum*) is a species endemic to northern China. Remaining numbers are probably in the range of 1 000 –

5 000 individuals. The species is now restricted to five or six disjunct areas, of which four are reserves: three in Shanxi and one in Hebei; a fifth population was recently found close to Beijing (McGowan, 1994). The species suffered a major decline historically owing to widespread deforestation of mountains within its geographic range. The species is classified as endangered under the revised IUCN Red List. Considering its minuscule distributional area, small population size, and endemism within China, the Chinese government has already listed the species on the "First-class Protected Birds List."

To predict the distribution of this species, we obtained base geographic data coverages from ESRI (1996). In all, 23 themes were included: snow cover (low and high); annual, January, and July precipitation (low and high); annual, January, and July temperature (low and high); land use; geography; solar radiation (low and high); human population; soils; vegetation; morphological structure; and frost-free period. We converted themes into ASCII raster grids using ArcView, and sent them to the BSW facility in San Diego to be mounted as base coverages. Covering all of China, these base coverages are now publicly available for other applications.

Searching the scientific literature (Salvadori, 1895; Peters, 1940; Cheng, 1979; Schauensee, 1984; Cheng, 1987) available to us, and with the kind collaboration of BirdLife International, we obtained 13 unique occurrence points for the species. Textual geographic references were then converted to latitude-longitude pairs by direct consultation of maps, resulting in the following sets of coordinates: 111.42°E 37.75°N, 111.48°E 37.90°N, 111.53°E 37.58°N, 111.57°E 38.70°N, 112.00°E 38.93°N, 112.20°E 39.43°N, 112.30°E 39.02°N, 112.50°E 37.92°N, 114.93°E 40.83°N, 114.99°E 39.85°N, 115.00°E 40.00°N, 115.33°E 40.00°N, and 116.50°E 40.25°N.

To perform the actual modeling, on the BSW facility, we chose China at base data, pasted in the geographic coordinates of the occurrence points at biological data upload, eliminated coverages not desired for analysis in modify data, and selected GARP and convergence criteria of 0.025 at make model. Once the model was built, the predicted map was displayed, and clicking at accuracy showed the predicting accuracy. Clicking overlays, we selected ARCgrid and again made a model, which produced an ASCII raster grid. This grid was copied, and saved in Word97 as an ASCII text file with line breaks.

In ArcView GIS (version 3.1), we imported the ASCII grid, which permitted further visualization of the prediction, including overlay with political geographic coverages. To further refine the prediction, we used published range maps to reduce the predicted distribution to areas likely inhabited (other areas are at times predicted because conditions appropriate for populations can exist in areas outside of the actual geographic distribution owing to historical factors such as limited colonization ability). ArcView's tabulate areas function was used to count pixels and predicted presence.

Given an average pixel size of 373.74 km^2 in the Chinese coverages used, we predicted the raw, unadjusted distributional area of the Brown Eared Pheasant in China to 376 730 km^2 (1 008 cells). However, taking into account the biogeographic limitations of the species (Peterson *et al.*, 1999b), its predicted distributional area was reduced to 174 910 km^2 (468 cells), which constitutes the species' potential distributional area. Finally, using U.S. Geological Survey Land use/and cover classifications based on AVHRR satellite imagery available at the EROS Data Center Website, we reduced our prediction to natural habitats within the species' potential distributional area, which totaled but 11 960 km^2 (32 cells) in seven isolated sites, one in Beijing, two in Hebei Province, and four in Shanxi Province.

While the most recent range estimates for the species suggest that it is now limited to about 13 600 km^2, our predicted area is 11 960 km^2. The two independent estimates are remarkably close; the difference could spring from the timing of the estimates and different versions of vegetation maps used. We used an updated vegetation map to refine our predictions. Because of deforestation, suitable habitat area for the species is decreasing.

3 Discussion

The approach described above opens exciting new opportunities for studying geographic distribution of species. The approach is based on contrasts between characteristics of known occurrence points and those of the landscape of the region in general. The general approach offers several important features: (1) tools are low-impact and fast, allowing interactive applications to be developed; (2) data formats are open, and can be integrated with custom scripts, permitting development of applications that bridge the gaps between "data," "analysis," and "visualization;" and (3) the modeling procedure is scale-independent, making possible applications at almost any spatial scale.

Several sources of potential error do exist in the modeling procedure. First, distribution will often be overpredicted (i.e. predicted area too large) because of omission of critical limiting ecological dimensions from the analysis. Although distributional predictions appear to stabilize with relatively small numbers of ecological dimensions, such errors can be detected only through addition of more ecological dimensions to the analysis, or via procedures such as jackknifing of the inclusion of dimensions to detect instability (Peterson *et al*., 1999a). Additional error in predictions commonly springs from historical influences on distribution (Peterson *et al*., 1999b): for example, a given species may not occur on a mountain range not for lack of appropriate conditions, but rather because intervening lowland habitats prevented its ever having reached that range. An effective solution to this complication involves limiting predicted areas to those biotic regions or geographic units from which the species has actually been recorded, providing that sampling has been sufficiently intensive as to make absence of such records reliable. Correction for these sources of error is thus feasible, if models and predictions are interpreted carefully, and with consideration of possible biases and confusions, yielding distributional predictions that are highly believable.

Potential applications of this approach of ecological niche modeling and distributional prediction are numerous, including the following:

①Prediction of distribution of rare and poorly known species, including species so rare that localization of populations is difficult without inferential approaches.

②Prediction of areas of potential distribution for rare and endangered species, permitting design of strategies for reintroduction of species to natural areas.

③Evaluation of niche dimensions of single species or multiple species for use in evolutionary and comparative studies of evolutionary change in ecological dimensions.

④Distributional prediction for suites of species of interest for conservation concern, allowing development of strategies for protected areas systems or evaluation of potential for negative environmental impacts.

⑤Use of ecological niche models for synthetic models predicting distributional shifts under scenarios of global climate change, or species invasions of presently uninhabited areas.

Further exploration and application by investigators with diverse interests will undoubtedly add many more possible applications to the list.

Acknowledgements Many thanks to David Stockwell for his masterful programming and development of the Biodiversity Species Workshop facility, to Miguel Ortega Huerta for expert technical assistance, and to Mike Crosby of BirdLife International for generously sharing distributional data.

References

Bojorquez-Tapia, Izuara I L, Ezcurra E, 1995. Identifying conservation priorities in Mexico through geographic information systems and modeling[J]. *Ecol. Appl.*, **5**:215 – 231.

Carpenter G, Gillison A N, Winter J, 1993. DOMAIN: A flexible modeling procedure for mapping potential distribution of plants and animal [J]. *Biodiv. and Conserv.*, **2**:667 – 680.

Cheng T H, 1979. Fauna Sinica Aves, Vol. 2: Anseriformes[M]. Beijing: Science Press.

Cheng T H, 1987. A synopsis of the avifauna of China[M] Beijing: Science Press

Daniels R J R, Hegde M, Joshi N V *et al*, 1991 Assigning conservation value: A case study from India[J]. *Conserv. Biol.*, **5**:464 – 475

Harrison J A, Martinez P, 1995. Measurement and mapping of avian diversity in southern Africa: Implications for conservation planning [J] *Ibis*, **137**:410 – 417.

Kiester A R, Scott J M, Csuti B *et al*, 1996. Conservation prioritization using GAP data[J]. *Conserv. Biol.*, **10**:1332 – 1342.

McGowan P J K, 1994. Family Phasianidae (pheasants and partridges) [A]. In: Josep del Hoyo, Andrew Elliott, Jordi Sargatal. Handbook of the birds of the world[M]. Barcelona: Lynx Edicions.

Nix H A, 1986. A biogeographic analysis of Australian elapid snakes [A]. In: Atlas of Australian Elapid Snakes[M] Canberra: Bureau of Flora and Fauna. 4 – 15.

Peters J L, 1940. Check-list of birds of the world, Vol. I – XVI[M]. Cambridge: Harvard University Press.

Peterson A T, Navarro-Siguenza A G, Benitez-Diaz H, 1998. The need for continued scientific collecting: a geographic analysis of Mexican bird specimens[J]. *Ibis*, **140**:288 – 294.

Peterson A T, Cohoon K C, 1999a. Sensitivity of distributional prediction algorithms to geographic data completeness[J]. *Ecol. Modelling*, **117**:159 – 164.

Peterson A T, Soberon J, Sanchez-Cordero V, 1999b. Conservatism of ecological niches in evolutionary time[J] *Science*, **285**:1265 – 1267.

Peterson A T, Egbert S L, Sánchez-Cordero V *et al*, 2000. Geographic analysis of conservation priorities using distributional modelling and complementarity: endemic birds and mammals in Veracruz, Mexico [J]. *Biological Conservation*, in press.

Russell-Smith J, Bowman D M J S, 1992. Conservation of monsoon rainforest isolates in the Northern Territory, Australia[J] *Biol. Conserv.*, **59**:51 – 63.

Salvadori T, 1895. Catalogue in the British Museum. Vol. V-XXVII[M]. London: Taylor and Francis Press

Schauensee R M D, 1984. The birds of China[M]. Washington: Smithsonian Institution Press.

Sperduto M B, Congalton R G, 1996. Predicting rare orchid (small whorled pogonia) habitat using GIS[J] *Photogr. Eng. & Remote Sens*, **62**:1269 – 1279.

Stockwell D R B, Noble I R, 1991. Induction of sets of rules from animal distribution data: A robust and informative method of data analysis [J]. *Math Comp. Simul.*, **32**:249 – 254.

Stockwell D R B, 1993. Bayesian learning system for rapid expert system development[J]. *Expert Syst Appl*, **6**:137 – 147.

Walker P A, Cocks K D, 1991. A procedure for modeling a disjoint environmental envelope for a plant or animal species[J] *Global Ecol. Biogeogr. Letters*, **1**:108 – 118.

一种利用推论性模型预测陆地脊椎动物分布的技术

Townse., A

陈国君　A. Townsend Peterson

（美国堪萨斯大学自然历史博物馆暨生态与进化生物学系　*Lawrence KS*　66045 – 2454）

摘要：对用于预测物种地理分布的新技术进行了叙述。这种预测包括以下步骤：建立地理基础资料库（database）、收集被预测物种的地理分布点以及使用生物多样性的附件——Genetic Algorthm for Rule-set Prediction 来建立被预测物种的生态位模型。为了能清楚地理解这一预测过程，以濒危物种褐马鸡（*Crossoptilon mantchuricum*）为例进行了叙述。这项新技术是用于生态学、生物地理学、分类学，保护生物学研究而对生物地理分布进行评估的有用工具。

关键词：地理信息系统；GARP；分布；生态位　陆地脊椎动物．地理分布．推论性模型

中图分类号：Q958.2　**文献标识码**：A　**文章编号**：0254 – 5853(2000)03 – 0231 – 07